Federated Access to Cyber Observables for Detection of Targeted Attacks

Michael Atighetchi, John Griffith, Ian Emmons, David Mankins
Raytheon BBN Technologies
Cambridge, MA
{matighet, jgriffit, iemmons, dm}@bbn.com

Richard Guidorizzi
Defense Advanced Research Projects Agency (DARPA)
Arlington, VA
richard.guidorizzi@darpa.mil

***Abstract* — Current DoD enterprise networks routinely face targeted cyber attacks, and even though attack-related information is recorded in various places, this information is often left unexamined until after attacker objectives have been achieved. This is especially true for large networks consisting of continuously changing networked devices, including laptops, servers, printers, IP phones, and more. This paper describes the design of Gestalt, a next-generation cyber information management platform that simplifies access to cyber event data stored in the nooks and crannies of a distributed enterprise. The ready and secure access to cyber information provided by Gestalt is a key enabler for a new set of techniques that can detect and mitigate targeted cyber attacks within hours instead of months. Current state-of-the-art approaches to automated and operator assisted cyber defense are ill-suited to counter targeted cyber attacks because these technologies (1) focus only on aggregated one-dimensional features across multiple devices, (2) do not provide the required coverage over all networked devices and observables accessible on those devices, and (3) lack the expressiveness and deeper semantic backing required to detect targeted attacks across a sea of low-level observables. Gestalt provides innovations in (1) automatically discovering devices and useful data sources in the enterprise (beyond simple IP connectivity), (2) maintaining a metadata index of devices and observable information (even of devices without schemas and connectors), and (3) transparently decomposing and federating semantic graph queries to devices (rather than extracting and aggregating information in a central store), and integrating the results back into a well-defined ontology.**

Keywords: cyber security, federated data access, Semantic Web, ontology, middleware

I. INTRODUCTION

Today, to detect the very specialized attacks typically launched against Department of Defense (DoD) information technology (IT) networks, cyber defenders must access and analyze information derived from a wide range of sources such as log files, operating systems and user-space executables, databases of various formats, device configurations, directory structures, communications paths, file and message headers, etc. Cyber defenders must employ a number of system-specific specialized tools to collect the information from each of the systems in the network required for analysis of a suspected attack. They must then analyze each information type and a manually cross-correlate the identified events. This manual process of data collection, correlation, and analysis is far too labor-intensive to keep pace with increasing attack frequency and sophistication.

Under DARPA's Integrated Cyber Analysis System (ICAS) [1] program, we are implementing *Gestalt*, a next-generation federated access solution which automates and simplifies cyber information management. The Gestalt system eliminates tedious manual inspection by providing access to all data sources on the network via a federated query interface. Using a new Cyber Defense Language, a single query can access data residing on multiple devices, across disparate device types and data formats, and return the query results in a semantically integrated and immediately useful format.

Gestalt allows the cyber defender to focus on the forensic data itself by abstracting away the actual methods and techniques required to access that forensic data. Through its Semantic Query Decomposition capabilities, Gestalt infers the types of data sources that can be used to satisfy a given query, and identifies where instances of those data source types can be found on the network. Next, it dispatches native queries to the device containing each data-source instance to process the request. The results are semantically integrated and returned to the cyber defender. Gestalt provides a single interface to the cyber defender, dramatically improving their effectiveness and allowing them to focus their time and expertise on forensic analysis of the results of their search queries, rather than on the laborious process of data collection and processing.

Gestalt gives the cyber defender the ability to create complex, multi-stage queries. For example, the single query:

> *Find all machines on which a user opened attachments and outbound connections to target machines within 5 minutes.*

can be combined with the subsequent query:

> *Find internal machines that received an inbound connection and created a new listening socket within 5 minutes.*

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
OCT 2014		**00-00-2014 to 00-00-2014**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Federated Access to Cyber Observables for Detection of Targeted Attacks	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Raytheon BBN Technologies,10 Moulton Street,Cambridge,MA,02138**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Military Communications Conference (MILCOM 2014), Baltimore, MD, October 6 - 8, 2014.

14. ABSTRACT
Current DoD enterprise networks routinely face tar-geted cyber attacks, and even though attack-related information is recorded in various places, this information is often left unex-amined until after attacker objectives have been achieved. This is especially true for large networks consisting of continuously changing networked devices, including laptops, servers, printers, IP phones, and more. This paper describes the design of Gestalt, a next-generation cyber information management platform that simplifies access to cyber event data stored in the nooks and crannies of a distributed enterprise. The ready and secure access to cyber information provided by Gestalt is a key enabler for a new set of techniques that can detect and mitigate targeted cyber attacks within hours instead of months. Current state-of-the-art approaches to automated and operator assisted cyber defense are ill-suited to counter targeted cyber attacks because these technol-ogies (1) focus only on aggregated one-dimensional features across multiple devices, (2) do not provide the required coverage over all networked devices and observables accessible on those devices, and (3) lack the expressiveness and deeper semantic backing required to detect targeted attacks across a sea of low-level observables. Gestalt provides innovations in (1) automati-cally discovering devices and useful data sources in the enterprise (beyond simple IP connectivity), (2) maintaining a metadata in-dex of devices and observable information (even of devices with-out schemas and connectors), and (3) transparently decomposing and federating semantic graph queries to devices (rather than extracting and aggregating information in a central store), and integrating the results back into a well-defined ontology.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**7**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

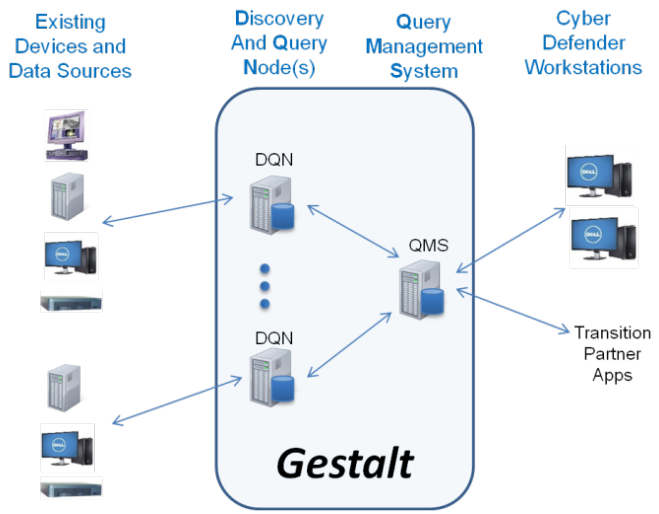

Fig. 1. The Gestalt system architecture minimizes network load while allowing cyber defenders to easily access information across the enterprise.

Such queries report intricate activity patterns in the network regardless of device type, host OS, or where logs reside and how they are accessed. This capability is achieved through the innovative application of BBN's Asio [2] semantic federated query technology and the Web Ontology Language (OWL) [3].

Gestalt's "leave the data in place" approach is in stark contrast to that of typical Security Information and Event Management (SIEM) [4] systems, which limit expressiveness to univariate statistics and centralize raw observables, thereby creating a high-value attack target and single point of failure. In addition to providing unprecedented ease of data access to the cyber defender, Gestalt imposes a minimal incremental impact to the subject network by utilizing a distributed, secure architecture, which pushes query processing and data manipulation to system nodes that reside in close proximity to where the data natively resides. Fig. 1 illustrates the Gestalt system architecture, highlighting the two primary system components: the Discovery and Query Nodes (DQNs) and the Query Management Service (QMS). The DQNs provide the interface between the devices on the network and the Gestalt system. The DQNs use standards-compliant protocols and definitions, including the Simple Network Management Protocol (SNMP) [5] and the Distributed Management Task Force (DMTF) Web Services for Management (WS-MAN) [6] technologies, to communicate with individual devices and to catalog the data sources each manages. The DQNs also utilize an intelligent information extraction capability for automatically suggesting mappings from text found in audit logs and web pages to concepts expressed in the Cyber Defense Language. Multiple DQNs are deployed within an enterprise to distribute query processing and device discovery across multiple devices and to distribute the network and processing load. The QMS maintains and applies the mappings between the data sources and the base ontology and performs the initial dispatching of queries to the DQNs. Cyber defenders access Gestalt through a rich web-based user interface that assists them with formulating queries, presents the results as processing is completed, and maintains a history of results and a "cyber defender's blog" to help them as they conduct investigations.

II. Motivating Example

Gestalt provides an advanced cyber event monitoring capability giving cyber defenders ready access to a wide range of cyber attack indicators across a diverse set of devices. At the highest level, Gestalt supports two main operational modes:

Continuous Monitoring: Cyber defenders specify semantic queries that Gestalt executes continuously to report back suspicious events that warrant further investigation. The queries are formulated over Cyber Defense Language (CDL), an ontology that represents concepts such as endpoints, flows, names, and checksums. Queries are then formed in SPARQL [7], a vendor-neutral graph query language.

Incident Response: Upon receiving external notifications of suspicious behaviors leaving the monitored network, e.g., expressed through minimal packet capture information such as the following

<source IP, destination IP, destination Port, timestamp, Rationale: "This traffic was going to a server that is known for Command and Control activities">

Based on information provided by any of these sources, the cyber defenders start an investigation to (1) confirm the event actually occurred, (2) confirm the event represents an attack, (3) identify the risk of the attack, (4) identify the source(s) of the attack, and (5) identify the device(s) affected by the attack.

To highlight the capabilities Gestalt is designed to provide in terms of targeted attack identification, let's consider two motivating examples that highlight complexities found in diverse and changing environments and provide context for the rich space of possible queries issued by highly skilled cyber defenders to find targeted cyber attacks.

The first example involves a targeted attack that (1) initially compromised a Windows 7 laptop via a spear phishing email containing a PDF attachment exploiting a zero day vulnerability in Adobe Reader (see APT1 [8] for examples), (2) established a toehold on the laptop by installing a custom backdoor that communicates with command and control servers using a number of encoding channels, e.g., DNS or Twitter, (3) proceeded to escalate privileges both locally and within the network to spread to other machines, including printers and servers containing sensitive information, and (4) exfiltrated sensitive information through a number of encoding channels.

The second example involves IP phones as an example of embedded network devices that have become pervasive within the DoD enterprise. These phone systems are often managed by a centralized server that stores information in a SQL database. The phones themselves can be configured to allow SSH access and provide device specific information via a local administrative interface offered by an embedded web server. A recent targeted-attack against IP phones [9] allowed arbitrary code execution on a large number of phones, at which point it was easy to turn the phone into an active attacker-controlled listening device or use the phone as an exfiltration router by establishing a H.323 connection to an attacker-controlled modem over the digital phone network mimicking a phone call, effectively eluding any digital intrusion detection capabilities.

Gestalt gives cyber defenders the ability to detect such attacks by virtue of having access to fine-grained cyber observa-

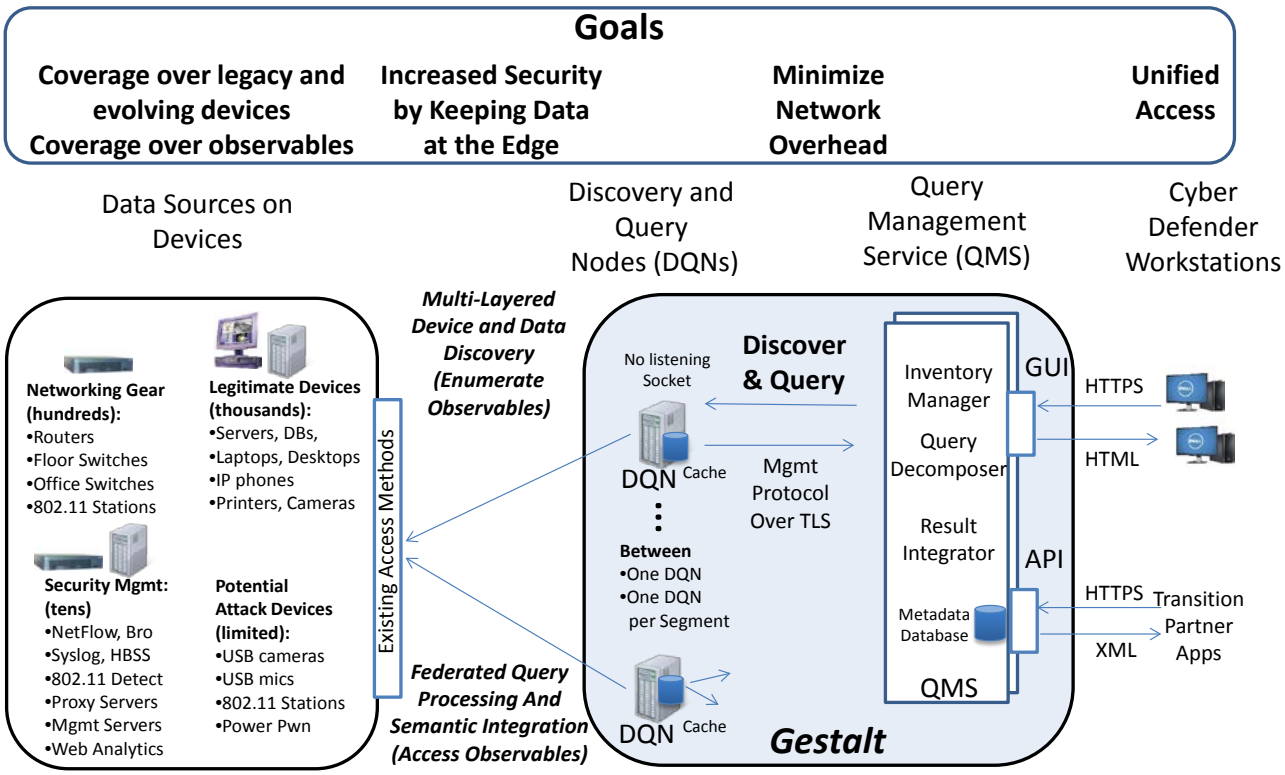

Fig. 2. Gestalt provides unified access to both legacy and evolving devices using a well-defined cyber ontology, while minimizing network overhead and ensuring a high level of security through its distributed architecture design.

bles including (1) process state, open file and socket state, and memory footprints to detect execution of 0-day exploits, (2) detailed information on outbound flows for common protocols, e.g., DNS (see [10] for tunneling attacks) and HTTP, (3) detailed information on internal flows, e.g., to spot Cisco HSRP attacks [11] allowing end systems to become trusted routers, (4) changes to application-specific configuration databases, e.g., plugin changes Firefox settings (about:config).

III. Related Work

The work presented in this paper relates to federated semantic database systems, cyber event monitoring, network monitoring, and stream processing/big data platforms.

The combination of all available cyber observables can be understood as a federated database system [12] with a large heterogeneity of access protocols and representation formats. Semantic integration of disparate data sources has been studied extensively in the Semantic Web community, including [13][14][15][16]. Gestalt uses best practices in ontology management, by separating data source ontologies from domain ontologies and reusing existing foundational ontologies [17], e.g., for expressing time [18] and events [19].

A number of commercially available solutions exist in the SIEM and cyber monitoring product space today, including ArcSight [4] and the Host Based Security System (HBSS) [20]. In contrast to Gestalt, these systems generally focus on keeping statistical summaries of events over time, aggregated and reported along various dimensions. While Gestalt provides detailed access to the current system state, SIEMs provide extended summary information at lower granularity.

A number of solutions exist for network and grid monitoring [21] , including Ganglia [22], Nagios [23], and Zabbix [24]. These systems specialize on performance monitoring and provide operators with dash-board views on the current availability state of the overall network system. Gestalt, in contrast, aims to provide visibility into detailed system state to detect loss of integrity (corrupted processes) and confidentiality (outbound exfiltration flows).

Finally, a number of big data platforms exist for distributed processing of information. Splunk [25][26] is a well-known instance of a big data processing capability that makes it easy for cyber defenders to establish correlations between disconnected pieces of text information through a specialized query language. Unlike Gestalt, Splunk is based on an information model that requires raw observables to be aggregated in a big data database before they can be queried.

IV. System Goals

The Gestalt system was designed with security in mind from the beginning, and in addition places significant focus on operating in contested network environments. Fig. 2 illustrates the Gestalt architecture, highlighting the four key system goals, namely; (1) unified access to all data source information in the IT environment through the introduction of a well-defined language and interface, (2) minimized network load through a distributed architecture that scales to large and complex IT environments, (3) keeping data at the edge by providing federated access to raw observables, and (4) coverage over both legacy and evolving devices and coverage over observables through the strategic combination of common access methods

for device-level interaction, and intelligent information extraction for handling new and evolving devices and data sources.

A. Unified Access

The overall effectiveness of the cyber defender is significantly improved by allowing a cyber defender to quickly and intuitively assemble information he or she needs to complete an investigation or identify effects of targeted cyber-attacks without getting distracted by the low-level processes involved in accessing, parsing, and transforming raw data. The result is that with more time to dedicate to the forensic analysis of the attacks rather than on the process of gathering data for that analysis, more complex investigations will be performed in less time with fewer resources. A side benefit is that as Gestalt handles these low-level data access and parsing details for cyber defenders, they require less training on the myriad of systems that make up an IT environment, allowing them to focus more on the conduct of forensics investigations.

Current systems provide cyber defenders with some aspects of what Gestalt provides, however they suffer from several critical limitations, primarily 1) the limited expressiveness of the query language they support, 2) an inability to easily accommodate new data source types, and 3) dependencies on a centralized data repository. A key component of the Gestalt system that enables the goal of Unified Access is a unified language and interface that provides integrated access to all data source information in the IT environment. Gestalt uses a semantic web cyber ontology to build the Cyber Defense Language with a rich graph of actors, devices, services, and observables, that are temporally connected by multi-dimensional relationships (e.g., accessed, modified, deleted, connected, started, stopped, transmitted) and temporally indexed. Gestalt also provides automated means for mapping existing data source into CDL, thereby reducing the amount of time required to provide access to existing data.

B. Minimized Network Overhead

Current systems provide cyber defenders with a composite picture only after creating a central repository where selected samples of the actual data are mapped and indexed. This leads to critical issues around volume, processing and timeliness. Building and maintaining a central repository consumes an increasing amount of the IT environment's network and processing resources. Restricting those resources results in a loss of data fidelity and exposes the IT environment to attack. Finally, it is highly unlikely that any individual data item would actually be used in a forensic examination – the resources consumed to copy, process, and store that data item are essentially wasted, as it is impossible to determine a priori what data items may eventually be needed. In contrast, Gestalt imposes a minimal incremental impact to the network due to its distributed, secure architecture that pushes query processing and data manipulation to system nodes.

C. Increased Security by Keeping Data at the Edge

Gestalt is a modular design that supports small infrastructures as well as very large and complex environments. The query function is designed to perform Gestalt related processing on or as close to the end-systems as the IT organization desires. Working with the raw data in its original location provides the highest level of data fidelity while simultaneously reducing the resources required to transmit, store and process it to the minimum level possible. The DQN can be deployed beyond a firewall or a constrained bandwidth link to further minimize impacts on the network while providing full forensics investigatory support.

D. Coverage over Legacy and Evolving Devices and Data Sources

The IT environment constantly evolves, introducing new data sources and changing the format of existing ones. This is especially true in military networks where forces and systems are constantly being assembled and revised to meet evolving mission needs. Providing 100% coverage of the administered devices and systems within the enterprise is a goal of the Gestalt system. Today, SIEMs use proprietary, special purpose data source handlers, and therefore must provide and support a constant stream of new components. This forces vendors to focus on widely-deployed data sources and systems, and ignore devices and applications with small market potential. A related goal is to provide a system that is 'future proof' – that is, a system that retains its effectiveness without redesign or redevelopment as the enterprise evolves. Adaptability and extensibility are key attributes of Gestalt.

This is accomplished using a combination of data source inventory and intelligent information extraction to automatically identify data source locations and infer mappings for an evolving set of sources and types. Upon receiving a semantic query, the QMS decomposes the query into sub-queries that are sent to the DQNs for execution on the end-systems. Gestalt provides semantic alignment for a wide set of data sources, e.g., SQL databases, Common Information Model (CIM) objects, SNMP Management Information Bases (MIBs) and log files, in a way that promotes extensibility and code reuse. This allows queries to be run directly on devices such as IP phones, and return only result data back to the QMS, distributing the processing needed to keeping most of the data on the device and limiting network load.

V. Core Functionality

The functionality that Gestalt provides consists of two fundamental functions that together enable access to cyber observables. *Discovery*, performed by the DQNs, finds new data sources that can be linked into the unified access scheme. The DQNs collectively maintain a meta-data index about devices and data sources which can be replicated to the QMS. During *query processing*, the QMS receives queries from cyber defenders and federates the queries to DQNs for resolution. Results are communicated from the DQNs back to the QMS and presented back to the cyber defenders.

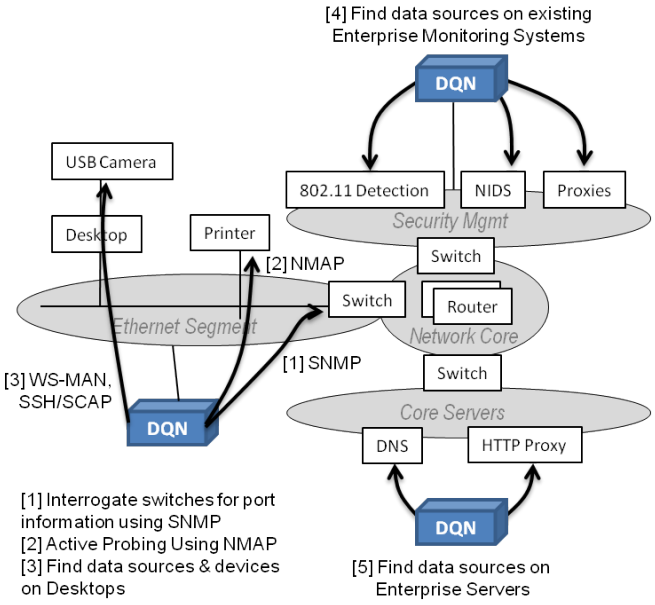

Fig. 3. Multi-Layered Discovery provides coverage over data sources while minimizing network load

A. *Discovery*

Fig. 3 shows the Gestalt multi-layered discovery process in more detail that produces a high-accuracy index of available devices and data sources. First, the DQN discovers and interrogate switches for a list of active MAC addresses and IP addresses found across network components (step 1). In addition, the SNMP-based discovery which minimizes network load can be augmented by active probing at the IP layer that involves running tools like NMAP to get a listing of live IPs and available services (step 2). The DQN then uses an existing SNMP, WS-MAN or remote shell service (identified in step 2) to identify data sources accessible on or through the device, such as USB devices and log files (step 3). DQNs use the same process on servers to identify log files already produced by existing enterprise monitoring systems, e.g. Bro & NetFlow, (step 4) and access logs produced by servers (step 5).

The integration of multiple corroborating pieces of information enables the DQNs to detect cases in which malware attempts to avoid detection in one dimension (e.g., by making a device unresponsive to NMAP requests) but not all (the malware eventually needs to exfiltrate data over the network). The output of the multi-layered discovery process is a metadata index, as shown in Fig. 4. The DQN manages the lifecycle of this index over the devices and data sources it is con-figured to monitor. Based on its configuration, the DQN can either replicate this information to the QMS (which enables the QMS to do a better job at query decomposition) or to keep it local (in which case queries might be unnecessarily decomposed and submitted to the DQNs by the QMS). Gestalt provides architectural flexibility to enable the proper operating point giving various competing constraints in the deployment environment.

The DQN also maintains a mapping between the data sources and the access wrappers used to execute queries on those data sources. The DQN ships with a set of pre-configured wrappers, and also contains a machine learning based wrapper

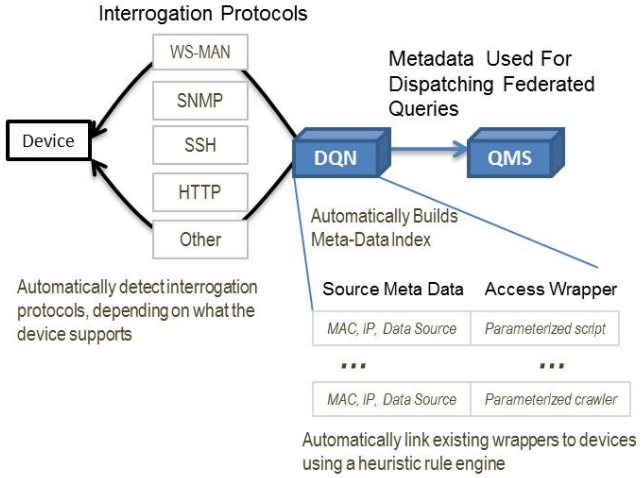

Fig. 4. The DQN automatically generates and index of devices and data sources and communicates the index to the QMS for the purpose of dispatching federated queries.

to infer mappings for previously unseen data sources that are similar to existing ones. The DQNs can also use a number of syntactic and semantic correlation heuristics to associate retrieved information with Gestalt event data types. For example, unstructured data from different web pages may present time and addressing information differently and in different locations within the page.

B. *Query Processing*

The QMS component serves as the primary interface through which cyber defenders use Gestalt to issue onetime as well as standing queries, expressed in the Cyber Defense Language, and obtain results, as shown on the right of Fig. 5 The QMS first performs authentication of cyber defenders, using X.509 certificates obtained through TLS 1.2 connections [27]. Upon successful authentication, requests are passed through an access control engine to authorize access based on attributes and roles. These roles include defined sets of DQNs that an individual is authorized to access, administrative actions permitted, and a System Security Officer role. Once query requests pass authorization, they enter a semantic query decomposition phase. Asio plays a major role in handling these queries, by first performing a Semantic Query Decomposition (SQD) and then dispatching the decomposed component-queries to appropriate DQNs or QMS-resident components that act as semantic bridges—bridging the semantic gap between raw, device resident data and the Cyber Defense Language in which query responses are expected. As shown in Fig. 5, a query issued by cyber defenders may be serviced by multiple data sources. Asio natively supports SQL databases, Web Services, and SPARQL endpoints. For Gestalt, DQN's adapter framework and access wrappers are designed to work as endpoints receiving and responding to query components dispatched by the QMS. The decomposition and dispatching are guided by the federated inventory and an index of devices and data sources that are created and maintained by the QMS and the DQNs.

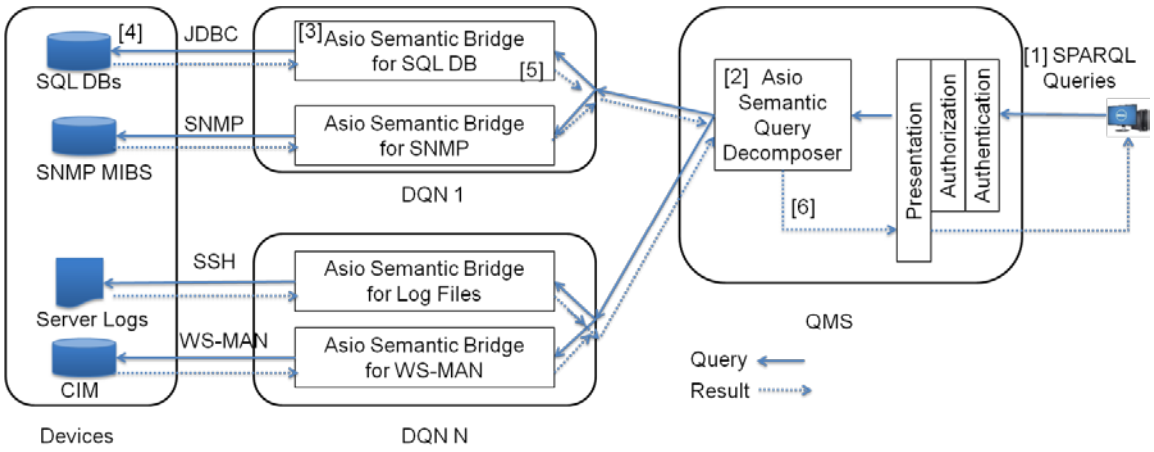

Fig. 5. The QMS authenticates and authorizes requests, then automatically decomposes and federates the queries across the DQNs.

The steps underlying how queries expressed in the Cyber Defense Language get mapped to endpoints that can retrieve the desired data from devices are best explained in terms of the example interaction shown in Fig. 5, as follows:

1. A cyber defender uses the QMS to issue a SPARQL query expressed in the Cyber Defense Language to the QMS.
2. The SQD automatically decomposes the query into a series of SPARQL sub-queries using mapping rules for each known data source. The mapping rules, expressed in the Semantic Web Rule Language (SWRL) [28], constitute Gestalt's device and data source inventory and index, are collectively maintained by both the DQNs and the QMS.
3. The semantic bridges translate the SPARQL sub-queries into the native query language of the underlying data source: SQL for Relational Databases (e.g., Host Based Security System (HBSS) or Cisco CallManager data), WS-MAN, SNMP MIBS, and log file patterns.
4. The sub-queries are executed on the target data sources. For instance, SQL *select* statements are executed via Java Database Connectivity (JDBC), extract scripts for log files are executed over SSH, or CIM objects are interrogated via WS-MAN.
5. The result sets are returned back up through the stack: first translated into RDF by the semantic bridges; then into the domain ontology via the mapping rules. This second translation is where the actual semantic alignment occurs.
6. The QMS returns the query result to the cyber defenders in form of web page updates.

This approach to data gathering and semantic alignment takes unified data access a step further than relational and Service Oriented Architecture (SOA) technology: Here all queries and results are expressed in the same internal representation ontology, allowing the cyber defenders to use a single domain vocabulary and expertise no matter what the source is. Note also that this unified access is achieved without disrupting the data sources – all of the applications that depend on those sources continue to work without change. For access to data sources, the DQNs having locally-persisted credentials necessary to access their monitored administrative domains, and only those domains. The credentials are stored in an encrypted key store on the DQN protected by a credential from the QMS.

This way, compromise of a single DQN would compromise administrative access only to a specific subset of the network, and defenders couldn't go around Gestalt to gain direct access to systems unless otherwise permitted by the administrative domain(s) themselves.

The benefit of Gestalt's Semantic-Web-inspired approach to data unification is that it can dramatically reduce the cost of software maintenance and enhancement, which are the major drivers when considering the Total Cost of Ownership for an end-user organization. Deploying Gestalt requires creating three configuration artifacts:

- **Domain** Ontology**:** In our case this is the Cyber Defense Language.
- **Source Ontology:** This is an OWL representation of the data source schema that enables the semantic bridge to map data between the native and Semantic Web formats. It is distinct from the domain ontology and for the most part is generated automatically with tools included in the Asio toolkit.
- **Mapping Rules:** For each data source, the integrator writes SWRL rules that map data expressed in the source ontology into the domain ontology.

Because each data source can be mapped into the domain with little regard for the mappings of other data sources, the integrator can focus on one source at a time. Thus, as the number of sources grows, the integration effort stays manageable. In contrast, relational and SOA integration complexity tends to grow as the square of the number of data sources and quickly becomes impractical. For exactly the same reasons, the marginal cost to add a new data source (or to modify the configuration when a data source changes) is minimal. This means that incremental expansion of the integration over time is not only possible, but is the favored integration approach and best supports evolving operational requirements. Finally, note that no extra work is required to support new queries, as is the case

with some integration approaches (such as a SOA). Because queries are translated from domain to source on the fly, even ad hoc queries are easily supported.

VI. High-Level Security Architecture

Since the security of the Gestalt system itself was a key design requirement, a number of mitigations exist for ensuring integrity, availability, and confidentiality of the DQN and QMS components and the data they make available. DQNs will be configured with the strongest access method available for each data source, e.g., SNMPv3, TLS 1.2 with approved ciphers, and be placed on isolated management networks. Authentication credentials required to access the data sources are persisted on the DQNs in a secure manner, e.g., in key stores with access controlled through SELinux policies. The DQNs will perform data filtering and sanitization on any data received from data sources, as the general assumption is that some of the data sources have been compromised. Interactions between the QMS and the DQNs are protected using mutually authenticated TLS 1.2 connections with approved ciphers.

VII. Conclusion and Next Steps

DoD enterprise environments are under constant attack by skilled adversaries that launch targeted attacks which can remain undetected for an extended period of time. Current commercial offerings focus on technologies sold in quantity to customers facing attacks identified as risks to a broad set of targets – both commercial and government. As such, these solutions do not provide the level of access to cyber observables that are needed to successfully mitigate targeted attacks.

This paper describes the architecture of a new advanced cyber information management system, called Gestalt, that is currently being implemented under the DARPA ICAS program to provide unified, secure, federated access to a wide range of cyber data sources. Gestalt's functionality consists of discovery and query processing performed over a set of federated DQNs that are centrally managed by a QMS. Gestalt's design directly supports 1) unified access to all data sources, 2) minimized network overhead, 3) increased security by keeping data at the edge, and 4) coverage over both legacy and evolving devices and coverage over observables.

Working from the basis presented in this paper, we are extending the existing implementation to support ingestion of BRO logs (for passive discovery) and walk the network of switches and routers using SNMP (for active discovery). Furthermore, we are currently implementing the management protocol between the QMS and DQNs, based on an asynchronous polling paradigm that provides both strong security and also works in contested network environments in which establishment of long-lived sessions is problematic. Finally, we will be evaluating the Gestalt design and implementation artifacts against an internally developed threat model (described using an attack tree) to establish security arguments for confidentiality, integrity, and availability of Gestalt. We also expect to engage transition partners in discussion about the cost benefit tradeoffs associated with performing a large number of granular measurements to spot suspicious behaviors in a real-time.

References

[1] DARPA, "Integrated Cyber Analysis System (ICAS) Homepage," 2014. [Online]. Available: http://www.darpa.mil/Our_Work/I2O/Programs/Integrated_Cyber_Analysis_System_%28ICAS%29.aspx.

[2] M. Fisher and M. Dean, "Semantic Query: Solving the Needs of a Net-Centric Data Sharing." Semantic Technology Conference, 23-May-2007.

[3] D. L. McGuinness, F. Van Harmelen, and others, "OWL web ontology language overview," *W3C Recomm.*, vol. 10, no. 2004–03, p. 10, 2004.

[4] D. Miller and B. Pearson, *Security information and event management (SIEM) implementation*. McGraw-Hill, 2011.

[5] W. Stallings, *SNMP, SNMPv2, SNMPv3, and RMON 1 and 2*. Addison-Wesley Longman Publishing Co., Inc., 1998.

[6] DMTF, "Web Services for Management (WS-MAN) Specification," DSP0226, 2010.

[7] J. Pérez, M. Arenas, and C. Gutierrez, "Semantics and Complexity of SPARQL," in *The Semantic Web-ISWC 2006*, Springer, 2006.

[8] M. I. Center, "APT1: Exposing one of China's cyber espionage units," Mandiant, Tech. Rep, 2013.

[9] Cisco, "Cisco Unified IP Phone Local Kernel System Call Input Validation Vulnerability," Cisco Security Advisory Advisory ID: cisco-sa-20130109-uipphone, Mar. 2013.

[10] D. Raman, B. De Sutter, B. Coppens, S. Volckaert, K. De Bosschere, P. Danhieux, and E. Van Buggenhout, "DNS tunneling for network penetration," in *Information Security and Cryptology–ICISC 2012*, Springer, 2013, pp. 65–77.

[11] E. Vyncke and C. Paggen, *LAN Switch Security*. Cisco Press, 2008.

[12] A. P. Sheth and J. A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Comput. Surv. CSUR*, vol. 22, no. 3, pp. 183–236, 1990.

[13] D. Kolas, "Query Rewriting for Semantic Web Information Integration," in *Proceedings of the Sixth International Workshop on Information Integration on the Web (IIWeb-07)*, Vancouver, Canada, 2007.

[14] M. Fisher, M. Dean, and G. Joiner, "Use of OWL and SWRL for Semantic Relational Database Translation," in *Proceedings of the Fourth OWLED Workshop on OWL: Experiences and Directions*, Washington, DC, 2008.

[15] "Revelytix Homepage." [Online]. Available: http://www.revelytix.com.

[16] C. Bizer and R. Cyganiak, "D2rq-lessons learned," in *W3C Workshop on RDF Access to Relational Databases*, 2007, p. 35.

[17] D. Kolas and T. Self, "Towards an Effective Methodology for Rapidly Developing Component-Based Domain Ontologies," in *Proceedings of the 2009 International Conference on Ontologies for the Intelligence Community*, Fairfax, VA, 2009.

[18] "OWL Time Ontology." [Online]. Available: http://www.w3.org/TR/owl-time.

[19] "Basic Formal Ontology (BFO) Homepage." [Online]. Available: http://www.ifomis.org/bfo.

[20] DISA, "The Host Based Security System," 2012. [Online]. Available: http://www.disa.mil/Services/Information-Assurance/HBS/HBSS.

[21] S. Zanikolas and R. Sakellariou, "A taxonomy of grid monitoring systems," *Future Gener. Comput. Syst.*, vol. 21, no. 1, pp. 163–188, 2005.

[22] M. L. Massie, B. N. Chun, and D. E. Culler, "The ganglia distributed monitoring system: design, implementation, and experience," *Parallel Comput.*, vol. 30, no. 7, pp. 817–840, 2004.

[23] W. Barth, *Nagios: System-und Netzwerk-Monitoring*. No Starch Press, 2008.

[24] R. Olups, *Zabbix 1.8 network monitoring*. Packt Publishing, 2010.

[25] J. Stearley, S. Corwell, and K. Lord, "Bridging the gaps: joining information sources with Splunk," in *Proceedings of the Workshop on Managing systems via log analysis and machine learning techniques*, 2010.

[26] Splunk.com, "Splunk for Security – Supporting a Big Data Approach for Security Intelligence," 2014.

[27] M. Atighetchi, N. Soule, P. Pal, J. Loyall, A. Sinclair, and R. Grant, "Safe configuration of TLS connections," in *Communications and Network Security (CNS), 2013 IEEE Conference on*, 2013, pp. 415–422.

[28] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, M. Dean, and others, "SWRL: A semantic web rule language combining OWL and RuleML," *W3C Memb. Submiss.*, vol. 21, p. 79, 2004.